Exploring Usage of Digital Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online

Gooding, Paul
University College London
paul.gooding.10@ucl.ac.uk

Long Paper

URL: dhwriter.org/paper/1383240494.251.310.5.html

Exploring Usage of Digital Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online

- Gooding,Paul
  University College London
  paul.gooding.10@ucl.ac.uk

Summary

This paper presents the findings of a case study of Welsh Newspapers Online. It demonstrates, through detailed content web log analysis, that while users are deeply engaged with digital newspaper collections, they tend not to browse newspapers. This confirms suspicions that users interact with digital newspapers online differently to the physical objects.

1. Introduction

This paper will present the findings from a case study undertaken with the support of the National Library of Wales. It utilises web log analysis to discover more aboutt he use and users of Welsh Newpspaers Online (WNO),[1] a digitised newspaper archive which currently contains around 420,000 digitised newspapers from, and relating to, Wales. Web log analysis has previously been undertaken to analyse online user behaviour, with utility for websites *[1] [2]*, e-journals *[3] [4] [5]*, and digital resources *[6] [7]*. To date, however, it has not been used to analyse the use of digitised newspaper archives. This paper presents the findings from a detailed analysis of content logs from a period of three months, starting from the launch of WNO in March 2014, and provides an important empirical study into how users are interacting with digitised newspapers. In doing so, it helps to illuminate some existing debates about the impact of digitisation upon reading and scholarship.

These debates have focused on how digitisation fo newspapers has changed the way users engage with these materials. Mussell, for instance, has noted that article-level representation foregrounds the partial textual transcript, even though the physical article is actually one textual component operating among many others. This means that users of a digitised newspaper may lose the original context of the material by approaching it in a different way online *[8]*. Brake additionally notes that "digital representations of nineteenth-century copy denaturalizes it and transforms the reader... into a *user* who sees the content inextricably embedded in the matrix of the newspaper pages" *[9]*. These issues have been expressed in concerns that this will have a sever impact on reading behaviour *[10] [11] [12]*. This paper demonstrates that, while user engagement remains extremely high, it is evident that the nature of this engagement has been unavoidably transformed by digitisation.

2. Methodology

In order to undertake this analysis, the National Library of Wales IT team supplied a complete set of content logs, collected for a period of three months from the launch of the collection in March 2013. These logs specifically recorded interactions with content discover and viewing mechanisms on the WNO website. As a result, they are not a complete log of the user's journey. Instead, the logs track a number of important basic behaviours: searchers undertaken by users on the website (search queries); occasions where users have browsed, filtered, or otherwise interacted with search results (search results queries); and instances where users have viewed newspaper pages (content queries). The weblogs record each interaction with the website as a single line of plain text, recorded automatically on the website servers. The following example demonstrates the format of a single content query:

2013-06-02T12:26:50+01:00 51a5c97c3c8d3 llgc-id:3036868 llgc-id:3039814 llgc-id:3037695 Aberystwyth Observer 21 September 1872 [2] ART40

The elements are, in order: date and time of interaction; unique user ID; server IDS for website content; title of newspaper viewed; date of newspaper edition; page number viewed; article number on the page viewed. Additionally, search queries contain a field with the user’s search terms, and search results queries include a field which shows the nature of the user’s interaction. In total, there were over 300,000 weblogs, which therefore provide a rich source of information about user interactions with the content of WNO.

This data allows several metrics to be analysed: most viewed newspaper titles; most viewed years; most viewed page numbers; average number of pageviews per visit; and percentage of pageviews involving search, search results or content queries. To achieve this, the investigator heavily post-processed the data in Excel. All non-relevant fields were stripped from the spreadsheet, then a column was added which automatically populated pageview numbers for each unique user ID, from 1 for the first page viewed by a user, to X for the last. Search, browser and content queries were also changed to numerical values of 1,2, and 3 respectively. This allowed the data to be processed into appropriate graphical representations. In particular, it allowed large-scale analysis of the complete dataset, in order to uncover overall patterns and trends in graphical form.

3. Findings

The content logs have provided a wealth of fascinating data on the behaviour and interests of users of Welsh Newspapers Online. We will discuss some of the most important findings in more depth below.

The following chart shows the findings from the large-scale analysis of the content logs, demonstrating the percentage of queries by type at each pageview. From this chart we can see which broad categories of user behaviour were most common for any given pageview:

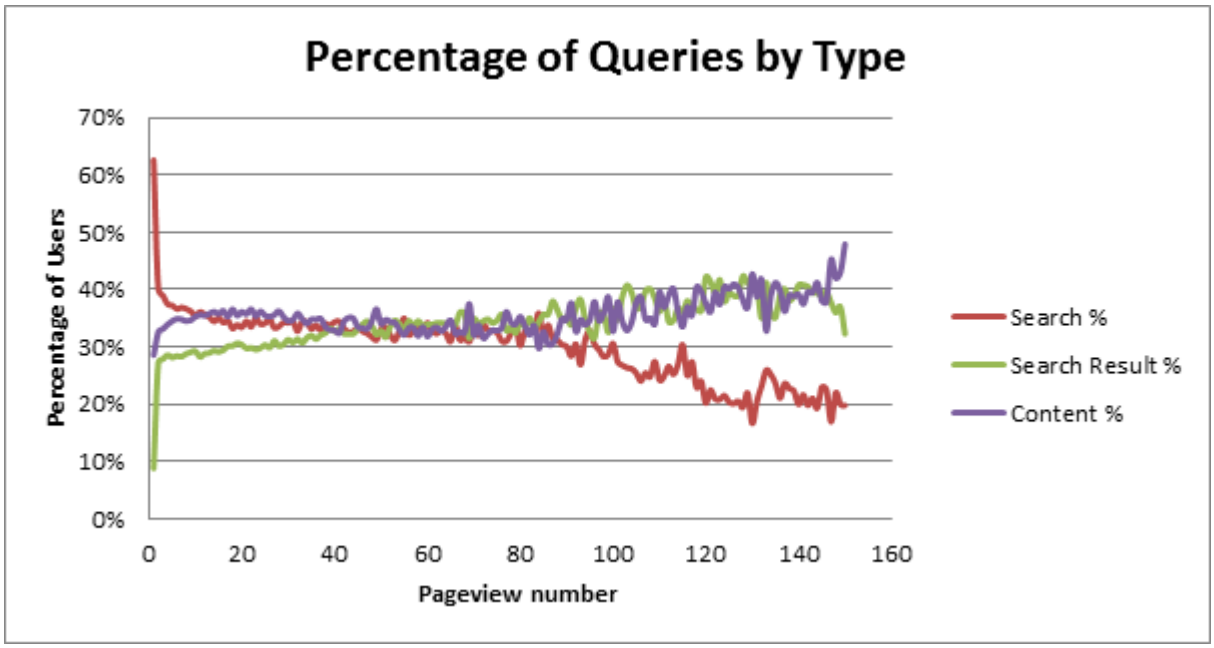

Fig. 1:

The chart demonstrates that content queries remain an important part of user activity, regardless of how long visitors spend on the website. In fact, once visitors view over 100 pages, they seem to view increasingly large amounts of content. By contrast, the longer a user spends on the resource, the less likely they are to be engaged in search activities. Instead, we see that search result queries replace search queries in importance.

This tells us that the importance of search as a discovery mechanism becomes less important with time spent on a particular visit: the majority of users begin by searching the collection, and then slowly but steadily move away from searching to view content or to browse search results. However, the data suggests that digital reproductions of newspapers do effect how users engage with them, in one particular way: while they browse the website, extensively, we can see that they do not browse through newspaper content. The following chart shows that users primarily tend to view the title page but, in general, the further into a newspaper the page is,

the fewer times it will be read. This data should be contextualised: the average number of pages per newspaper is around 6.3. However, the drop in views begins from the second page, which strongly suggests that users are unlikely to browse through pages:

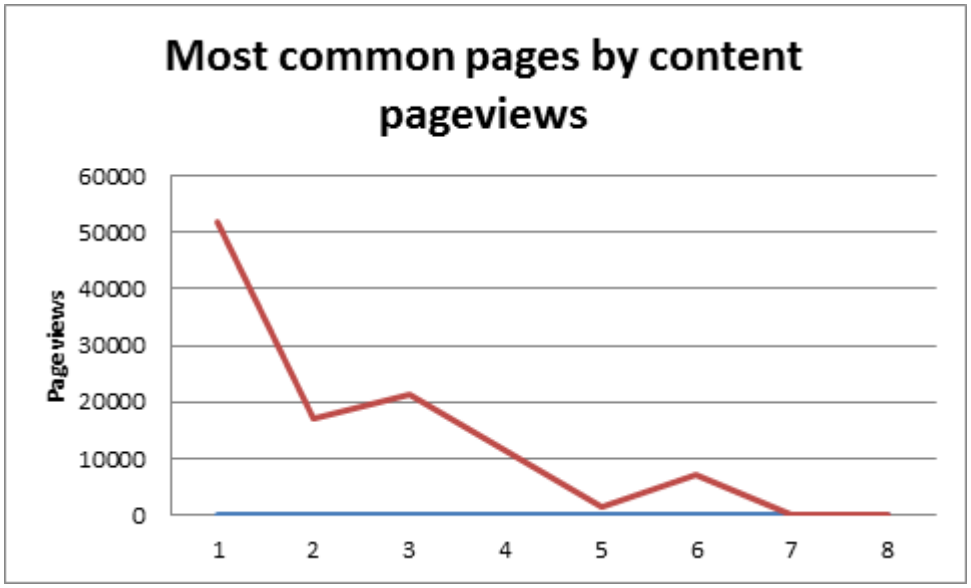

Fig. 2:

In this respect, the findings seem to confirm the fears that digitised texts interfere with the deep, ordered reading behaviour attributed to printed materials. While there are valid concerns about the way in which inline reading may distance users from the original context of newspaper material, we must also recognise that users certainly are deeply engaged with digitised newspaper archives: it is the type of engagement that has changed. This highlights two important conclusions: the side-lining of browsing makes serendipitous discoveries less likely, and means that search functionality, OCR quality, and website design are as important as the availability of content in ensuring that users are able to discover content; and that existing remediations of newspapers are some way from recreating the reading behaviour that has been attributed to the physical text *[13] [14]*.

4. Conclusion

This case study demonstrates that content log analysis can greatly enrich studies of the impact of digitised collections. They provide empirical evidence of user behaviour, which allows a granular, nuanced understanding of how researchers interact with digitised content online. We can therefore learn a great deal more about this understudied area. The findings also confirm, though, a commonly recorded problem with content logs; they provide a strong statistical base for analysis, but they cannot provide an insight into why users behave how they do. As such, an analysis which relies solely on content logs is necessarily incomplete. This paper will finish by proposing that web log analysis can still play an important role in analysing and understanding usage of digital resources, but that it should not be used in isolation.

---

[1] http://papuraunewyddcymru.llgc.org.uk/en/home?

References

1. Nicholas, D. et al., 2000. Evaluating consumer website logs: a case study of The Times/The Sunday Times website. Journal of Information Science, 26(6), pp.399–411.
2. Almind, T.C. & Ingwersen, P., 1997. Infometric Analyses on the World Wide Web: Methodological Approaches to 'Webometrics'. Journal of Documentation, 53(4).
3. Institute for the Future, 2002. E-Journal user study: report of web log data mining, Available at: http://ejust.stanford.edu/logdata.html [Accessed March 23, 2013].

4. Nicholas, D., Jamali, H.R. & Huntington, P., 2005. The use and users of scholarly e-journals: a review of log analysis studies. Aslib Proceedings, 57(6), pp.554–571.
5. Yu, L. & Apps, A., 2000. Studying e-journal user behaviour using log files: the experience of SuperJournal. Library and Information Science Research, 22(3), pp.311–338.
6. Warwick, C. et al., 2008. If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities. Literary and Linguistic Computing, 23(1), pp.85–102.
7. Meyer, E.T. et al., 2009. Usage and Impact Study of JISC-Funded Phase 1 Digitisation Projects & the Toolkit for the Impact of Digitised Scholarly Resources (TIDSR), Oxford: Oxford Internet Institute, University of Oxford. Available at: http://microsites.oii.ox.ac.uk/tidsr/sites/microsites.oii.ox.ac.uk.tidsr/files/TIDSR_FinalReport_20July2009.pdf [Accessed May 9, 2013].
8. Mussell, J., 2012. The Nineteenth-Century Press in the Digital Age, Basingstoke: Palgrave MacMillan.
9. Brake, L., 2012. Half Full and Half Empty. Journal of Victorian Culture, 17(2), pp.222–229.
10. Brake, L., 2012. Half Full and Half Empty. Journal of Victorian Culture, 17(2), pp.222–229.
11. Birkets, S., 1994. The Gutenberg elegies: the fate of reading in an electronic age, New York: Ballentine Books.
12. Deegan, M. & Sutherland, K., 2009. Transferred Illusions: Digital Technology and the Forms of Print, Ashgate Publishing. p.49.
13. Baker, N., 2002. Do we want to keep our newspapers? In Do we want to keep our newspapers? London: Office for Humanities Computing, pp. 19–34.
14. Birkerts, S. (1994). The Gutenberg Elegies: The Fate of Reading in an Electronic Age. Boston: Faber and Faber.